Computational Linguistics in the Netherlands Journal 3 (2013) 148-173 Submitted 06/2013; Published 12/2013

A graph-based approach for implicit discourse relations

Yannick Versley VERSLEY@SFS.UNI-TUEBINGEN.DE

SFB 833, Univ. Tübingen
Nauklerstr. 35
Germany

Abstract

Recognizing and classifying implicit discourse relations is a challenging task since hardly any strong indicators exist, and a variety of weak indicators has to be harnessed to yield evidence for a particular discourse relation or another. Most current approaches rely on a combination of shallow, surface-based features and rather specialized hand-crafted features, with a considerable gap in between which is partly due to the sheer complexity of combining evidence from different levels of linguistic description.

As a way to avoid both the shallowness of word-based representations and the lack of coverage of specialized linguistic features, we use a graph-based representation of discourse segments, which allows for a more abstract (and hence generalizable) notion of syntactic (and partially of semantic) structure, we propose an approach to use a graph-structured representation of discourse units in order to improve the classification of implicit discourse relations.

We validate this approach using implicit discourse relation data from the TüBa-D/Z treebank, providing an extended discussion and error analysis that looks at the impact of the graph-based representation on the different kinds of discourse relations.

The empirical evaluation shows that our graph-based approach not only provides a suitable representation for the linguistic factors that are needed in disambiguating discourse relations, but also improves results over a strong state-of-the-art baseline by more accurately identifying *Temporal*, *Comparison* and (for the German data) *Reporting* discourse relations.

1. Introduction

Discourse relations are semantic or rhetorical relations that hold between textual spans. They capture essential structural and semantic/pragmatic aspects of the coherence of a text. Besides anaphora and referential structure, discourse relations are a key ingredient in understanding a text beyond single clauses or sentences. The automatic recognition of discourse relations is therefore an important task; approaches to the solution of this problem range from heuristic approaches that use reliable indicators (Marcu, 2000) to modern machine learning approaches such as Lin et al. (2009) that apply broad shallow features in cases without such indicators.

Especially on *implicit discourse relations*, where no discourse connective could provide a reliable indication, broad, shallow features such as bigrams or word pairs conceivably lack the precision that would be needed to improve disambiguation results beyond a certain level. Conversely, hand-crafted linguistic features allow one to encode certain relevant aspects, but they have often limited coverage. Encoding detailed linguistic information in a structured representation, as in the work presented here, allows us to bridge this divide and potentially find a golden middle between linguistic precision and broad applicability.

Using a corpus of German with discourse annotation, the TüBa-D/Z treebank (Telljohann et al., 2009; Versley and Gastel, 2013) and a novel representation of discourse units as graph structures, we argue that such an approach is suitable to overcome the shallowness of a word-based representation and the non-specificity or lack of coverage of specialized linguistic features. We provide results for using this approach on a German corpus of discourse relations using a structure mining approach and an approach using support vector machines and convolution kernels.

The rest of this article is organized as follows: Section 2 takes a closer look at the concept of discourse relations; Section 3 gives an overview of existing approaches to use structure in automatic classification. Section 4 gives an overview of the corpus we used, and Section 5 describes the feature-based and graph-based representations that are used in this study. Section 6 describes the setup that is used to evaluate different variants of the system, and provides details on each experiment.

2. Discourse relations

The meaning of a text does not just consist in (purely) the propositions of its constituting sentences, but is conventionally understood as including the inferences that readers add so that the text grows into a coherent whole (implicatures). In this process of understanding the text, they are guided by the assumption that constituent parts of the text are in a specific relation to some other part of the text. In the following example (1), the clause *John pushed him* is taken to be the explanation for the asserted event *Peter fell* and we can express this insight by postulating a *causal* discourse relation between the two clauses.

(1) Peter fell. John pushed him.

Many theories of discourse assume that discourse relations both build a hierarchy of longer and longer parts of a text (complex discourse units), and can also hold between these larger parts. Rhetorical Structure Theory (Mann and Thompson, 1988, RST), in particular, assumes that the whole discourse consists of a hierarchy of (elementary, then complex) discourse units in the form of a projective tree. In other cases, such as Segmented Discourse Representation Theory (Asher, 1993, SDRT), a non-projective hierarchy is assumed (Afantenos and Asher, 2010), and in the assumptions underlying the Discourse GraphBank (Wolf and Gibson, 2005), the idea of hierarchy is more or less abandoned.

While the idea that discourse segments are linked by discourse relations belonging to one or multiple types seems to be universally accepted among different theories, the question of discourse units forming a hierarchy, and the exact constraints that hold on that hierarchy, has received different, mutually incompatible, answers within linguistic theories of discourse, and consequently in different existing resources, which means that any computational approach for taking into account not just labeling of discourse relation but also the identification of their structure across a text will be tied to one particular theoretical framework.

Because of this, we exclude the identification of structure from the considerations of this article. Instead, the focus is on the problem of identifying which discourse relation holds between the two spans of text, more specifically on those where no discourse connective gives an overt clue that can be used to identify this relation (*implicit discourse relations*).

In cases where a discourse connective links two discourse segments (*explicit discourse relation*), both the arguments and the realized discourse relation are actually unambiguous. For example, in "[*Peter despises Mary*] because [*she stole his yoghurt*]" the discourse relation is unambiguously signaled.

In other cases, a connective can be ambiguous, as in the case of German '*nachdem*' (as/after/since). *Nachdem* can signal multiple types of discourse relations (e.g. purely temporal or temporal and causal), as in (2):[1]

(2) [Nachdem sowohl das Verwaltungsgericht als auch das Oberverwaltungsgericht das Verbot bestätigt hatten,] [rief die NPD am Freitag nachmittag das Bundesverwaltungsgericht an].
[*After both the Administrative Court and the Higher Administrative Court had confirmed the interdiction,*]
[*the NPD appealed to the Federal Administrative Court.*] *(Temporal+cause)*

[1] TüBa-D/Z corpus, sentence 7462

Another type of discourse relations are *implicit discourse relations*, which can occur between neighbouring spans of text without any discourse connective signaling them:[2]

(3) [Mittlerweile ist das jedoch selbstverständlich]
[Die gemeinsame Arbeit hilft, den anderen zu verstehen.]
[*In the meantime, this has become a matter of course*] (implied:since) *(Explanation)*
[*The common work helps to appreciate the other.*]

In existing research, most earlier work such as Marcu (2000) or Soricut and Marcu (2003), as well as later Pitler and Nenkova (2009), exploit discourse connectives and syntactic context for the identification of discourse relations, while work such as Haddow (2005) and Miltsakaki et al. (2005) uses linguistic indicators such as tense and modality to improve results on ambiguous connectives. In the realm of ambiguous discourse connectives, Versley (2011a) claims that additional semantic and structural information – classes of adverbials, or the detection of contast pairs (such as *hot–cold*) can help improving the classification accuracy in such cases.

In the case of implicit discourse relations, the absence of overt clues makes successful disambiguation considerably harder, and a combination of weak linguistic indicators and world knowledge is needed for successful disambiguation. Sporleder and Lascarides (2008) use positional and morphological features, as well as subsequences of words, lemmas or POS tags to disambiguate implicit relations in a reannotated subset of the RST discourse treebank (Carlson et al., 2003). Sporleder and Lascarides also show that (despite the corpus size of about 1000 examples) actual annotated relations are more useful than artificial examples derived from non-ambiguous explicit discourse relations.

Research using the implicit discourse relations annotated in the second release of the Penn Discourse Treebank (Prasad et al., 2008) shows a focus on shallow features: Pitler et al. (2009) find that the most important feature in their work on implicit discourse relations are word pairs. Lin et al. (2009) identify production rules from the constituent parse, as well as word pairs, to be the most important features in the system, with dependency triples not being useful as a features, and information from surrounding (gold-standard) discourse relations having only a minimal impact. Park and Cardie (2012) provide a summary and synthesis of implicit relation labeling on the Penn Discourse Treebank: they identify context-free productions as the most effective feature across all relations, with a feature identifying common verb classes (as encoded in VerbNet, Kipper et al., 2000), as well as semantic classes and sentiment.

Some existing research deals with inferring both the structure and the labeling of discourse relations in a text, such as Hernault et al. (2010), Feng and Hirst (2012) on the RST Discourse Treebank (Carlson et al., 2003) or Muller et al. (2012) on the SDRT-based AnnoDis corpus (Péry-Woodley et al., 2011). In these cases, identification of discourse connectives and shallow indicators are again the mainstay of classification, while Feng et al. propose the semantic similarity between words of the two clauses as an indicator.

In summary, most existing research on automatic processing discourse relations treats individual clauses as bags of words, with only verbs and negators given special status. In order to use of the information in the clauses more effectively, it would be desirable to involve their internal structure in the classification process.

3. Structured classification

In the realm of machine learning approaches that are suitable to taking structure into account, one can distinguish structured-input learning techniques, which take a complex structure (molecular graphs of proteins, syntactic structures of a parsed sentence, local patterns in an image and their neighbourhood relations) and provide a simple output in terms of a yes/no label, or a category from

[2] TüBa-D/Z corpus, sentence 448

a fixed set, from structured-output learning techniques, where a complex structure is built from parts (such as in part-of-speech tagging, or parsing).

In both cases, local parts of the overall structure (such as a part-of-speech tag and the word it belongs to) can be labeled as belonging to a particular class (in the case of structured-input classification) or as being indicative of a good or bad output structure (in the case of a hypothesis in structured-output classification).

For our problem, the classification of graphs representing clauses into discourse relation categories, we care more about using larger parts of the structure in our classification (as is typical of structured-input classification), while not being concerned about the interdependencies between different discourse relation tokens (which would be typical in the case of structured-output classifiation such as RST-style building of discourse structure).

3.1 Graph classification in computational linguistics

All approaches for structured-input learning fall into one of three groups: *linearization* approaches, which decompose a structure into parts that can be presented to a linear classifier as a binary feature, *structure boosting* approaches, which determine the set of included substructures as an integral part of the learning task, and *kernel-based methods* which use dynamic programming for computing the dot product in an implied vector space of substructures. Kernel-based methods on trees have been used very widely, for example in the reranking of parse trees on a small treebank (Collins and Duffy, 2002) and for answer ranking in a question answering system (Moschitti and Quarteroni, 2011). A boosting algorithm for trees has been used by Kudo et al. (2004) for a sentiment task (classifying reviews into positive/negative instances). Arora et al. (2010) use subgraph features in a linearization-based approach to sentiment classification.

3.2 Partial tree kernels

Kernel-based methods such as support vector machines use a learning approach that does not need an explicit feature representation, but instead only depends on the kernel product $\kappa(x_1, x_2)$ between two learning instances. Tree kernels, or convolution kernels more generally, allow it to use an implicit feature space of arbitrarily-sized substructures without the efficiency problems that an explicit expansion of a structure into substructures would entail. Explicit expansion of structures into a list of substructures would yield an exponentially growing feature space, since there are exponentially many substructures. In contrast, dynamic programming methods allow it to use an implicit expansion that takes into account arbitrarily large substructures while only using a polynomial amount of time.

The *partial tree kernel* (Moschitti, 2006) uses an implicit feature space that is derived from all trees that have a subsequence of sibling nodes (but the same hierarchical relations between nodes).

As an example, consider a tree (written in bracketed notation) `(A (B (C D) (E)))`, which has the partial trees `(A)`, `(A (B))`, `(A (B (E)))`, `(B (C))`, `(B (C D))`, `(E)` etc. and a tree `(A (B (D) (E)))` , which has the partial trees `(A)`, `(A (B))`, `(A (B (E)))`, `(B (D))`, `(B (D))`, `(E)`, but not `(B (C))`, `(C)` or `(A (B (C)))` etc. The common substructures `(A (B))`, `(B)` or `(E)` would all contribute to the kernel product between these trees.

3.3 Linearization-based approaches

Approaches based on linearization decompose the tree (or graph) into a vector of fragment indicators as an explicit feature representation. As mentioned above, this explicit feature representation can be less efficient, which is why they are usually paired with techniques that filter for frequent subgraphs – such as gSpan (Yan and Han, 2002) – and/or techniques for generic feature selection. In addition, boosting approaches integrate learning and feature selection (Kudo et al., 2004), whereas recent approaches for tree kernel linearization (Severyn and Moschitti, 2013) first perform kernel-based

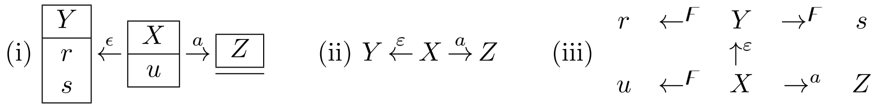

Figure 1: Example Feature-Node Graph (i), its backbone (ii), and its expansion (iii)

learning followed by a step that extracts the most important substructures from the support vector classifier that has been learned in order to use these as features in a linear classifier.

Both the kernel-based approach and linearization-based approaches are combineable with nominal (i.e., non-structured) features, but the general simplicity of a straightforward linearization approach makes it the best option for the starting point in our exploration of structure-based modeling of clauses, as reported in the next sections.

3.4 Feature-node graphs

When we want to represent subparts of a clause and their features, extracting features locally and then using a single (global) feature vector loses the information on each part. A better way to make use of the (implicit) information which property relates to which part of an argument span is to represent them in *feature-node graphs*, which we introduce here. This formalism also allows us to take into account more structure than n-grams (which are limited to relatively shallow information) or dependency triples (which would be too sparse in the case of typical discourse corpora).[3]

Formally, a feature-node graph consists of a set V of vertices with labels $L_V : V \to L$, a set of edges $E \subseteq V \times V$ with labels $L_E : E \to L$, with the addition of a set $F : V \to \mathcal{P}(L)$ that assigns to each vertex a set of *feature* labels.

The *backbone* of a feature-node graph is simply the labeled directed graph (V, L_V, E, L_E), without any features.

The *expansion* of a feature-node graph is the labeled directed graph (V', L'_V, E', L'_E) built by expanding the set of nodes to $V' = V \uplus \{(v,l) \in V \times L | l \in F(v)\}$ with labels $L'_V(v) = L_V(v)$ for all $v \in V$ and $L'_V((v,l)) = l$ for all $v \in V, l \in F(v)$ and correspondingly adding edges to get the complete set $E' = E \uplus \{(v,(v,l)) | l \in F(v)\}$, with a special symbol $\digamma$ for the labels of newly introduced edges, i.e. $L_E(v,(v,l)) = \digamma$.

Figure 1 gives an example of a feature-node graph with the vertices X, Y and Z with $F(X) = \{u\}$, $F(Y) = \{r,s\}$, and $F(Z) = \emptyset$, edges $E = \{(X,Y),(X,Z)\}$ and edge labels $L_E((X,Y)) = \varepsilon$, $L_E((X,Z)) = a$.

Representing desired information as features (instead of, e.g., using words, or POS tags, as the node labels in a dependency graph) is advantageous because that two feature-node graphs of similar structures will have a common substructure as long as the backbone of that structure is identical. In the case of words as node labels, any non-identical word would prevent the detection of the common substructure.

4. Discourse relations in the TüBa-D/Z corpus

In order to test our approach to discourse relation classification, we rely on a subcorpus of the TüBa-D/Z corpus that has been annotated with discourse relations. This subcorpus has received full annotation for all discourse relations, according to an annotation scheme described by Versley and Gastel (2013). This corpus contains 803 implicit discourse relations that are not marked by a connective (according to the criteria set forth by Pasch et al., 2003).

[3] For reasons of efficiency as well as learnability, the structures we use to represent each discourse unit are simpler and more compact than the annotated corpus data from which they are derived.

In the corpus, 41 texts have been annotated with a scheme postulating a hierarchical discourse structure that has elementary discourse units at its bottom end (described below), and larger discourse segments (*topic segments*) corresponding to a high-level communicative goal (*explain the history of the museum*) at its upper end, where a text is subdivided into multiple topic segments corresponding (usually) to the result of the author's high-level text organization.

At the lowest level, elementary discourse units (EDUs) are defined in syntactic terms as tensed clauses (matrix clauses and non-center-embedded subclauses), but also right-dislocations and non-restrictive relative clauses, as in the following example (4):[4]

(4) [Er sollte unseren heimischen Markt aufmischen,]
[das erste Produkt in deutschen Läden, das genmanipulierten Mais enthält.]
[*It was to stir up our domestic market,*]
[*the first product in German stores to contain GM corn.*] *(Restatement)*

In this example, the same situation (a chocolate bar containing genetically modified corn was put on the market) is described using two different aspects of the event (firstly, the intended effect, and secondly, background information about the chocolate bar).

Because the discourse structure is organized hierarchically, unmarked discourse relations can also occur between larger spans, such as in the *Commentary* relation of Example (5):[5]

(5) [$_\alpha$ Die gute Nachricht: Die Weltbevölkerung wächst inzwischen langsamer als in den vergangenen Jahrzehnten.| Die schlechte Nachricht: Erreicht wird diese Entlastung der ökologischen und sozialen Systeme nicht nur durch Fortschritte bei der Geburtenkontrolle,| sondern auch durch eine Sterblichkeitsrate, die zum erstenmal seit 40 Jahren wieder ansteigt. |. . .]

[$_\beta$ Diese Entwicklung zeigt nach Angaben von Lester Brown, einem der Autoren der Studie, das "Versagen unserer politischen Institutionen".]

[$_\alpha$ *The good news: The world population is growing more slowly than in past decades.| The bad news: This unburdening of ecological and social systems is not only achieved by progress in family planning,| but is also due to a mortality that growing again for the first time since 40 years. | . . .*]

[$_\beta$ *This development shows, according to Lester Brown, one of the study's authors, a "failure of our political institutions."*] *(Commentary)*

The *Commentary* relation in this example is typical for the progression from factual reporting in discourse segment α to a perspectivized opinion in segment β (which is attributed to *one of the study's authors*, but is also understood to be the opinion of the writer of the text at hand).

The relation labels in the annotation scheme are summarized in Table 1, and are grouped into the following broad categories:

- **Contingency** relations include those relations that would be termed *causal source of coherence* in the property scheme of Sanders et al. (1992). They include both causal relations in the narrow sense, but also conditionals ("*If I buy an umbrella, I will not get wet.*") and concession-like relations ("*Although he bought an umbrella, he still got wet.*")

- **Expansion** relations provide additional information to something introduced in one of the relation arguments. In the case of *Elaboration*, the statement is specified further by providing a redescription of the event (*Restatement*), an instance for the general phenomenon described in the segment (*Instance*), or some background information that can help understanding the plausibility or relevance of the first discourse segment (*Background*). *Interpretation* relations link a reported text segment to a summary or conclusion that is either purely factual (*Summary*) or contains claims that the author makes from their perspective (*Comment*).

[4] TüBa-D/Z, sentence 5736

[5] TüBa-D/Z sentences 8429ff.

Relation	# total	# implicit	% implicit	% relation
Contingency				
├ Causal				
│ ├ Result	133	88	66.2%	11.0%
│ └ Explanation	122	81	66.4%	10.1%
├ Conditional				
│ ├ Consequence	26	5	19.2%	0.6%
│ ├ Alternation	7	2	28.6%	0.2%
│ └ Condition	13	—	0.0%	—
└ Denial				
├ ConcessionC	60	9	15.0%	1.1%
├ Concession	34	5	14.7%	0.6%
└ Anti-Explanation	3	3	100.0%	0.4%
Expansion				
├ Elaboration				
│ ├ Restatement	149	140	94.0%	17.4%
│ ├ Instance	63	39	61.9%	4.9%
│ └ Background	119	109	91.6%	13.6%
├ Interpretation				
│ ├ Summary	2	1	50.0%	0.1%
│ └ Commentary	36	28	77.8%	3.5%
└ Continuative				
├ Continuation	89	71	79.8%	8.8%
└ Conjunction	45	1	2.2%	0.1%
Temporal				
├ Narration	127	70	55.1%	8.7%
└ Precondition	34	23	67.6%	2.9%
Comparison				
├ Parallel	55	23	41.8%	2.9%
└ Contrast	66	26	39.4%	3.2%
Reporting				
├ Attribution	67	67	100.0%	8.3%
└ Source	65	65	100.0%	8.1%

%implicit: proportion of relation instances that are implicit, rather than explicit. *% rel*: percentage of given relation among all implicit. About 10% of the implicit instances have multiple labels (e.g. *Result+Narration*).

Table 1: Frequencies of discourse relations in the corpus of Gastel et al. (2011)

- **Temporal** relations hold between events that are part of the same temporal sequence (*Narration*), or where the Arg1 of the relation is situated temporally in the post-state of the event of Arg2 (*Precondition*).
- **Comparison** relations center on two topical entities in Arg1 and Arg2 and compare them according to one particular property or attribute, focusing on the entities' similarity in that respect (*Parallel*) or their dissimilarity (*Contrast*).
- **Reporting** relations hold between a fact that is reported by one of the actors in the text and the event that constitutes the reporting. In the case of *Source*, the reported fact is seen to be asserted by the author (veridical). In the case of *Attribution*, the reported proposition is not necessarily portrayed as veridical, and the reporting event is seen as more central.

Among the most frequent unmarked relations are *Restatement* and *Background* from the Expansion/Elaboration group, which predominantly occur as implicit discourse relations, as well as *Result* and *Explanation*, which occur unmarked in about two thirds of the cases. In other cases, such as *Consequence*, *Concession* (is limited to cases of contraexpectation) and *ConcessionC* (which also includes more pragmatic concession relations), only a minority of relation instances is implicit whereas the majority is marked by an explicit connective.

Relations that are typically marked, such as *Contrast* – see Example (6) – or *Concession*/ *ConcessionC* – see Example (7) – often contain weak indicators for the occurring discourse relation, such as the opposition *policemen-demonstrators* in the first case, or the negation of a reference to Arg1 (“*this wish will not be fullfilled soon*”).

(6) [159 Polizisten wurden verletzt.]
[Zahlen über verletzte DemonstrantInnen liegen nicht vor.] *(Contrast)*
[*159 policemen were injured.*][*No data is available regarding injured demonstrators.*]

(7) [“Nun will ich endlich in Frieden leben.”]
[Dieser Wunsch Ahmet Zeki Okcuoglus wird so bald nicht in Erfüllung gehen.]
[*“Now I finally want to live in peace.”*] (implied: However,)
[*This wish of Ahmet Zeki Okcuoglu will not be fulfilled any time soon.*] *(ConcessionC)*

4.1 Annotation layers of TüBa-D/Z

The TüBa-D/Z is a multilayer corpus that contains the following annotation layers:

- A **word layer** containing the document’s tokens with associated part-of-speech, morphological, and lemma information.

- A **syntactic layer** containing a phrase structure parse for each sentence. The annotation scheme for phrase structure is aimed to capture theory-independent assumptions about the syntax of German while keeping as much information as possible within the projective phrase structure itself. The syntactic layer uses topological fields (Höhle, 1986) to organize the constituents occurring within a clause.

- A layer of **named entities** that represents the spans of named entities (which may or may not coincide with the spans of noun phrases) and their semantic class. The named entity annotation distinguishes between persons (PER), organizations (ORG), locations (LOC), geopolitical entities, which are territories with an elected or inherited governing entity (GPE), and other named entities such as works of art (OTHER).

  The spans of named entities usually coincide with the boundary of a noun phrase in the syntactic layer, but often this is not the case exactly, for example when the ‘canonical’ name does not include the determiner or a premodifier of the noun phrase. For example, in the noun phrase “[$_{NP}$ *Die Caritas*]”, only *Caritas* is part of the canonical name of the organization, yielding an annotation of “*Die* [$_{ORG}$ *Caritas*]” on the Named Entity layer.

- A layer with **coreference** information, which contains a pointer from a subsequent mention of an extratextual entity to its previous mention in the document, and which covers both anaphora (`anaphoric`) and definite descriptions and names (`coreferential`). The coreference layer also marks non-referring pronouns such as expletives (`expletive`) and inherent reflexives (`inherent_reflexive`). In case of a pronoun referring to an entity introduced by several noun phrases (ex.: *John met Mary at the station. Then <u>they</u> went home*), the `split_antecedent` relation is used.

While it is straightforward to use the word and the syntactic layer once the span of a discourse segment has been mapped to a sentential nonterminal node, named entities have to be mapped back to the syntax node to which they correspond.

5. Representing discourse segments

5.1 Feature-based representations

5.1.1 Linguistic features

We implemented a group of specialized linguistic features, which are inspired by those that were successfully used in related literature (Sporleder and Lascarides, 2008; Pitler et al., 2009; Versley, 2011a).

As implicit discourse relations can occur intra- as well as intersententially, the **topological relation** between the arguments is classified by syntactic embedding (if one argument is in the pre- or post-field of the other), or as one preceding, succeeding or embedding the other.

Several features reproduce simple **morphosyntactic properties**: One feature signals the presence of *negation* in either argument, either as a negating adverb (English *not*), determiners (*no*), or pronouns (*none*). A negated Arg1 would be tagged `1N+`, a non-negated one as `1N-`. *Tense and mood* of clauses in either argument are also incorporated as features (e.g. `1tense=t` for an Arg1 in pas(t) tense). The **head lemma(s)** of each argument, which is normally the main verb, is also included as a feature (e.g. `1Lverletzen` for the Arg1 of Example (6)).

We also mark the **semantic type of adjuncts** present in either relation argument, with categories for temporal, causal, or concessive adverbials, conjunctive focus adverbs (*also*, *as well*), and commentary adverbs (*doubtlessly*, *actually*, *probably* . . .). As an example, an Arg1 containing "*despite the cold*" would receive a feature `1adj_concessive`. Because this feature is based on a word list, it is imperfect in the sense that the list is incomplete, and some markers can be ambiguous with respect to their function.

The detection of **cotaxonomic relations** between words in both arguments uses the German wordnet GermaNet (Henrich and Hinrichs, 2010). Such pairs of contrasting lemmas, such as *hot-cold* or *policeman-demonstrator* commonly indicate a *parallel* or *contrast* relation. If two words share a common hyperonym (excluding the uppermost three levels of the noun hierarchy, which are not informative enough), feature values indicating the least-common-subsumer synset (such as *temperature adjective*) and up to two hyperonyms are added. For example, the pair *sagen* (to say) in Arg1 co-occurring with *erzählen* (to recount) in Arg2 would yield several features that include the synset IDs of the least common subsumer in GermaNet and its hyperonyms (e.g., `lcs_super_48092` — say something to someone, and `lcs_super_48077` — to utter).

A **sentiment** feature uses the lists of emotional words and of 'shifting' words (which invert the emotional value of the phrase) by Klenner et al. (2009) as well as the most reliable emotional words from Remus et al. (2010). The combination of emotional words and shifting words into a feature is similar to Pitler et al. (2009): according to the presence of positive- or negative-emotion words, each relation argument is tagged as `POS`, `NEG` or `AMB`. When a negator or shifting expression is present, a "`-NEG`" is added to the tag, yielding, e.g. "`1 pol NEG-NEG`" for an Arg1 phrase containing the words '*not bad*'.

Regarding the accuracy of these features, the identification of positive-sentiment or negative-sentiment items in arguments purely based on a word list (even together with shifting expressions) is relatively error-prone, since available resources mix actual sentiment terms with terms that have a positive or negative connotation within a certain domain (for example, a *weak* nuclear force in physics is not a sentimentally laden term, while a political figure having *weak* arguments would have a negative connotation). A more complete account of sentiment could be expected to be useful in all those cases where the question under discussion is an evaluative one.

Among the other features, the detection of cotaxonomic relations and the identification of semantic types of adjuncts are probably rather on the conservative side since they rely on hand-coded data rather than some kind of semi-supervised learning approach. (For a use as feature, in this case, limited coverage should be preferred over noise problems with extracted features).

5.1.2 Shallow features

As mentioned in Section 2, shallow lexical features empirically constitute a very important ingredient in the automatic classification of implicit (and ambiguous explicit) discourse relations, despite the fact that they lack most – semantic or structural – generalization capabilities. We implemented three groups of features that have been identified as important in the prior work of Sporleder and Lascarides (2008), Lin et al. (2009) and Pitler et al. (2009).

A first group of features captures (unigrams and) **bigrams** of words, lemmas, and part-of-speech tags. In this fashion, the bigram "*Zahlen über*" from Arg2 of Example (6) would be represented by word forms `2w_Zahlen_über`, lemmas `2l_Zahl_über` and POS tags `2p_NN_APPR`.[6]

Word pairs, i.e., pairs consisting of one word from each of the discourse relation arguments, have been identified as a very useful feature for the classification of implicit discourse relations in the Penn Discourse Treebank (Lin et al., 2009; Pitler et al., 2009), and, quite surprisingly, also for smaller datasets such as the discourse relations in the RST Discourse Treebank targeted by Feng and Hirst (2012) or the ambiguous connective dataset used by Versley (2011a).[7] Because of the morphological richness of German, we use lemma pairs across sentences; for Example (6), the lemma *Polizist* from Arg1 and the lemma *DemonstrantIn* from Arg1, among others, would be combined into a feature value `wp_Polizist_DemonstrantIn`.

Finally, **CFG productions** were used by Lin et al. (2009) to capture structural information, including parallelism. Context-free grammar expansions are extracted from the subtrees of the relation arguments and used as features by marking whether the corresponding rule type occurs only in one, or in both, arguments. In Example (6), the CFG rule 'PX → APPR NX' for prepositional phrases occurs in both arguments, yielding a feature "`pr B PX=APPR-NX`", whereas the preterminal rule "APPR → über" only occurs in Arg2 (yielding "`pr 2 APPR=über`").

5.2 Representing discourse units as graphs

In order to get a relatively general representation of discourse units in graph form, the **backbone** of the graph should be as simple and possible: in this fashion, commonalities are not hidden by variation in the node labels that is due to (incidental) variation of the syntactic form. In our approach, the **backbone** of the graph is built using nodes for a clause (`S`), and including children nodes for any clause adjuncts (`MOD`), verb arguments (`ARG`). In the case of relation arguments being in a (syntactic) matrix clause - subclause relationship (e.g. [$_{\text{Arg1}}$ *Peter wears his blue pullover,*] [$_{\text{Arg2}}$ *which he bought last year*]), the graph corresponding to the matrix clause receives a special node (`SUB-CL`, or `REL-CL` for relative clauses). This is universally the case for the explicit relations in the case of *nachdem*, but may also occur in the case of unmarked relations. For example, *Background* relations are frequently realized by relative clauses. Non-referring noun phrases (which are tagged as 'expletive' or 'inherent reflexive' in the referential layer of TüBa-D/Z), receive a node label `expletive` instead of `ARG`.

In each of the adjunct/argument nodes, we include **syntactic information** such as the category of the node (nominal/prepositional/adverbial phrase, e.g. `cat:NX` for a noun phrase), the topological field (cf. Höhle, 1986, e.g. `fd:MF` for a constituent occurring in the middle field) and, for clause arguments, the grammatical function (subject, accusative or dative object or predicative complement – e.g., `gf:OA` for the accusative object). Clause nodes contain features for tense and mood based on

[6] Sporleder and Lascarides (2008) use a Boosting classifier (BoosTexter) that can extract and use arbitrary-length subsequences from its training data. As our dataset is small enough that we do not expect a significant contribution from longer sequences, we approximate the sequence boosting by extracting unigrams and bigrams. As with the other shallow features, unigrams and bigrams are subject to the same supervised feature selection that is also applied to subgraph features.

[7] For an illustration of the differences in size, consider that the Penn Discourse Treebank contains about 20 000 implicit discourse relations in 2159 articles, and the RST Discourse Treebank contains a lower number of 385 documents; Sporleder and Lascarides used a sample of 1 051 annotated implicit relations which were derived from the RST Discourse Treebank but manually relabeled according to an SDRT-like annotation scheme.

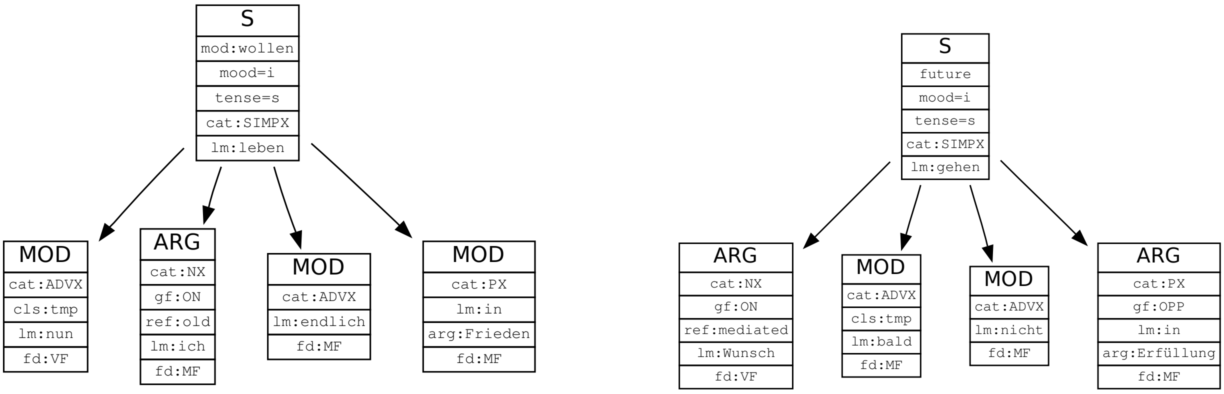

Figure 2: The complete graphs built from the implicit relation arguments "*Nun will ich endlich in Frieden leben.*" and "*Dieser Wunsch Ahmet Zeki Okcuoglus wird so bald nicht in Erfüllung gehen.*" – cf. ex. (7).

the main and auxiliary/modal verb(s) of that clause (e.g., `mood=i`, `tense=s` for an indicative/past clause).

In the realm of **semantic information**, we use the heuristics of Versley (2011a) to identify *semantic classes of adverbials*, in particular temporal, causal or concessive adverbials, conjunctive focus adverbs, and commentary adverbs. As the backbone of our graph structure abstracts from syntactic categories and only distinguishes adjuncts and arguments, it is possible to learn generalizations over different realizations of the same type of adjunct: for example, temporal adjuncts may be realized as a noun phrase (*next Monday*), a prepositional phrase (*in the last week*), an adverb (*later*), or a clause (*when Peter was ill*).

Noun phrase arguments are annotated with information pertaining to their **information status**, marking them either as *old* (if their referent has already been introduced), *mediated* (if a modifier – e.g. the genitive *John's* in *John's hat* – has been previously introduced), or *new* (if neither the phrase nor any of its modifiers has a previous mention). Additionally, we use a semantic categorization into persons (PER), organizations (ORG), locations (LOC), events (EVT) and other entities. In the case of named entities, this information is derived from the existing named entity annotation in the TüBa-D/Z treebank (by simply mapping the GPE label to LOC); for phrases with a nominal head, this information is derived using the heuristics of Versley (2006), which use information from GermaNet, semantic lexicons, and heuristics based on surface morphology. The identification of semantic classes has about 80-90% accuracy, lacking especially in cases of polysemy (e.g., process-product), but offering a precision over 90% for person instances. Clauses as well as arguments and adjuncts are annotated with their **semantic head**; prepositional phrases are, in addition, annotated with the semantic head of the preposition's argument (*in the next year*).

From the graph representations of relation arguments that are created in this step, frequent subgraphs are extracted. The subgraphs are then filtered based on thresholds which were chosen based on intuition. Considering the small size of the dataset, it would be useful to catch any subgraph that occurs recurringly, with some safety margin to allow for the fact that clauses may be (part of) an argument in multiple discourse relations, hence subgraphs must occur at least five times in either the Arg1 or Arg2 graph. Graphs also have (potentially) a greater-than-polynomial number of subgraphs, which is why we limit them to seven nodes. Because we want to capture the structure of the graph, rather than building arbitrary subsets of features, we could also want to require some of the subgraph nodes to be part of the backbone, or limit the number of nodes that are feature rather than backbone nodes (by default, to three). We investigate the sensitivity of the classification approach to parameter settings in the subgraph extraction in the following Section 6.

6. Experiments

For the 803 implicit discourse relations in the annotated subcorpus of TüBa-D/Z, we use a 10-fold cross-validation scheme where, successively, one tenth of the data is automatically labeled by a model from the remaining nine tenth of the data. Multiple relation labels are predicted by using binary classifiers (one-vs-all reduction) and using confidence values to choose one or several labels among those that have the most confident positive classification. In the case of multiple positive classifications (e.g., if *Reporting*, *Temporal* and *Expansion* all receive a positive classification), relations are only considered for the 'second' label if the most-confident label and the potential second label have been seen together in the training data (e.g. *Contingency* and *Temporal* can occur together, but *Reporting* will not be extended by a second relation label). In a second step, the coarse grained relation label (or labels) is extended up to the finest taxonomy level (e.g., an initial coarse-grained *Contingency* label is extended to *Contingency.Causal.Explanation*). In our experiments, we use SVMperf, an SVM implementation that is able to train classifiers optimized for performance on positive instances (Joachims, 2005).

6.1 Feature selection

Both the shallow features and the subgraph features are relatively noisy in the sense that only a minority of the actual feature values (i.e., a random subsequence of the lemmas in a clause, or a random pair of one word from one relation argument and one from the other) is indicative for the assignment of a discourse relation, whereas the rest of the feature values are not. The standard solution to this problem is to use *feature selection* and simply filter out all feature values that do not match a particular criterion such as frequency or correlation with the target labeling.

Unsupervised feature selection methods, such as setting a threshold on occurrence frequency, or on the entropy of a feature value as in (Versley, 2011a) do not make use of the target labeling.

In contrast, *supervised* methods for feature selection use the labeling information in the gold data to assess whether a feature value should be informative or not. Both Pitler et al. (2009) and Lin et al. (2009) and subsequent work use supervised feature selection techniques. Essentially, these methods use a correlation measure such as pointwise mutual information (in the case of Lin et al.) or the chi-square test between the presence of the feature value and each individual relation label, and take the highest score as the overall score for that feature value. In the actual classification, the training data is filtered so that it contains only the n feature values that have the highest overall score for the selection metric.

6.2 Experiment I: Feature-based representation

In a first experiment, we wanted to assess the relative utility of the different aspects of the feature-based representations. For this purpose, we consider each feature group (linguistic features *ling*, bigrams *bi*, word pairs *wp*, productions *pr*) by itself, then combinations between linguistic features and each group of shallow features (*ling+bi*, *ling+wp*, *ling+pr*). Finally, a combination of all feature groups (*all/nogr*).

For single groups of shallow features and combinations of linguistic and shallow features, we give the results for different thresholds for feature selection.

6.3 Experiment II: Graph-based representation

In a second experiment, we wanted to see which features of the clause representation are the most useful. We consider three variants:

- **Variant A** is the full graph, which includes syntactic information, semantic classes, and information status as well as lexicalization in the form of the semantic head.
- **Variant B** is the graph with the information status removed.
- **Variant C** contains the graph without any lexicalization or semantic information.

Since the classification infrastructure is a means for *using* representations of clauses, its usefulness crucially depends on having an informative clause representation. Considering these different variants allows us to see how sensitive the classification approach is to the richness or paucity of information.

6.4 Experiment III: Subgraph extraction versus tree kernels

As discussed in Section 3, there are multiple approaches to using a graph-based representation in learning and classification for discourse relations. Contrasting two learning approaches using the same representation helps us to understand the essential properties that a successful approach of learning with structured input representation would need.

6.5 Experiment IV: Varying the filtering criteria for subgraphs

As discussed in Section 3, we apply different thresholds in the extraction of subgraphs, which were initially based on intuitions regarding efficiency and performance. Looking at the system behaviour when one varies these parameters gives us a better idea of the nature of the graphs that contribute to improving the classification performance.

6.6 Experiment V: Comparative experiments for English

In order to test the portability of the graph-based approach to other languages, we use data from the Penn Discourse Treebank (Prasad et al., 2008). In order to stay relatively close to the setting for the smaller German corpus, we perform 10-fold crossvalidation on the validation set of the Penn Discourse Treebank (sections 00–02), instead of training on much more data and predicting the test set results only once. Our subset of the PDTB data contains 1958 relation tokens in total.

The annotation schemes of the Penn Discourse Treebank and of the TüBa-D/Z discourse annotation are relatively similar, with several exceptions. Firstly, the PDTB uses a separate mechanism for the annotation of attribution relations and treats them separately from the discourse relations, unlike the solution in TüBa-D/Z which treats them as discourse relations in the *Reporting* group. Secondly, concession relations (*Peter likes to travel but he gets seasick very quickly*) are part of the *Comparison* group in the PDTB while they are part of the *Contingency* group in TüBa-D/Z. Finally, event causation (*Peter stumbled, and fell on his nose*) receives both a temporal relation and a causal relation (e.g., *Narration+Result-cause*) in the TüBa-D/Z whereas only one relation is chosen in the Penn Discourse Treebank.

In the validation subset of the Penn Discourse Treebank, the most frequent group of relations is the *Expansion* relation (1025 occurrences), followed by *Contingency* (526 occurrences), *Comparison* (299 occurrences) and *Temporal* (111 occurrences).

For the test with the Penn Discourse Treebank data, we use a smaller set of linguistic features that are close to the feature set of Versley (2011b) and a version of the graph representation that contains the basic clause/argument/adjunct backbone, semantic heads, and function labels, but no semantic information or indicators of information status.

7. Results and discussion

7.1 Evaluation measures

In our study, the classification results in a complex prediction that can contain multiple labels, which in turn come from a taxonomy of discourse relations. As a result, a simple percent accuracy figure is insufficient as an evaluation statistic, and we need to consider both different granularities of relations and give partial credit when a relation with multiple labels (e.g. *Temporal+cause*) has been assigned a labeling with a subset/superset or overlapping categories (e.g., *Temporal*, or *Temporal+parallel*).

As a way to assign partial credit, we use a variant of accuracy that uses the **Dice score** to assign partial credit to relation tokens with an overlap between prediction and gold annotation.[8]

Given a set G of labels that constitute the labeling of a relation in the gold annotation, and a set S of labels that the system assigns as a labeling, the Dice score for one relation token is calculated as

$$\mathrm{Dice}_{\mathrm{tok}}(G_i, S_i) := \frac{2|G_i \cap S_i|}{|G_i| + |S_i|}$$

Where G_i and S_i are the relation labels assigned to one particular relation token i.

To give a few examples, an exact match would be scored as 1.0, whereas guessing a sub- or superset (e.g. only *Result* instead of *Result+Narration*) would give a contribution of 0.66 for that example, and overlapping predictions (*Result+Comparison* instead of *Result+Narration*) would get a partial credit of 0.5.

The Dice score for a system is simply the sum of the Dice scores of all relation tokens divided by the number of relation tokens in the dataset. Equivalently,

$$\mathrm{Dice}(G, S) := \frac{1}{N} \sum_i \mathrm{Dice}_{\mathrm{tok}}(G_i, S_i)$$

Because it averages over relation tokens, the Dice score is not very sensitive to the performance of a system on minority relations – indeed, the solution of always assigning the most frequent relation can give a Dice score that is very hard to beat. If one is interested in how well a system does across all relations, it is more informative to average over *relation types* instead of over *relation tokens*. Calculating the F_1 scores of each individual relation type (label) and then the average over all possible relation types yields the **Macroaveraged F-Score (MAFS)**.

Formally, we can write this as

$$\mathrm{MAFS}(G, S) := \frac{\sum_{\mathrm{rel} \in \mathcal{R}el} F_1(G_{\mathrm{rel}}, S_{\mathrm{rel}})}{|\mathcal{R}el|}$$

Because the macroaveraged F-score gives a (proportionally) larger importance to rare relations, it is also less stable with respect to small changes in the system response, whenever these pertain to less-frequent relation labels.

7.2 Quantitative results

Table 2 provides evaluation figures for different subsets of the presented features, using aggregate measures over relations both at the coarsest level (for implicit discourse relations, the five categories *Contingency*, *Expansion*, *Temporal*, *Comparison*, *Reporting*), and the finest level (which contains twenty-one relations in the case of implicit relations).

For each level of granularity, we can measure the quality of the classifier's predictions in terms of an average over relation tokens, the *Dice score*, which assigns partial credit for a relation token

[8] The Dice score is similar to a microaveraged F-measure. For instances with multiple labels, the Dice score assigns less importance than for those with only a single label, while microaveraged F-score would weight them proportionally to the number of labels.

	5 relations		21 relations		Cont	Expn	Temp	Comp	Rept
	Dice	MAFS	Dice	MAFS	F_1	F_1	F_1	F_1	F_1
Restatement	0.474	0.129	0.161	0.014	0.00	0.65	0.00	0.00	0.00
random	0.338	0.233	0.096	0.056	0.27	0.50	0.06	0.21	0.14
ling only	**0.540**	**0.396**	**0.274**	0.127	0.40	0.68	0.32	0.00	0.58
bi(500)	0.471	0.251	0.201	0.066	0.23	0.64	0.04	0.00	0.34
bi(1k)	0.513	0.283	0.245	0.082	0.32	0.65	0.00	0.00	0.44
bi(2k)	**0.517**	0.283	0.257	0.089	0.31	0.65	0.02	0.00	0.44
bi(5k)	0.516	**0.301**	0.260	0.098	0.40	0.65	0.00	0.00	0.45
wp(500)	0.469	0.284	0.207	0.067	0.28	0.63	0.12	0.00	0.40
wp(1k)	0.469	0.252	0.197	0.070	0.20	0.64	0.05	0.00	0.37
wp(2k)	**0.494**	**0.307**	0.198	0.084	0.42	0.65	0.02	0.05	0.40
wp(5k)	0.489	0.297	0.200	0.083	0.41	0.66	0.00	0.05	0.37
pr(500)	0.474	0.129	0.184	0.026	0.00	0.65	0.00	0.00	0.00
pr(1k)	**0.479**	0.142	0.192	0.036	0.04	0.65	0.00	0.00	0.03
pr(2k)	0.478	**0.185**	0.199	0.046	0.27	0.66	0.00	0.00	0.00
pr(5k)	0.478	0.154	0.192	0.034	0.12	0.65	0.00	0.00	0.00
ling+bi(5k)	0.545	0.399	**0.300**†	0.141	0.39	0.69	0.33	0.00	0.59
ling+wp(2k)	**0.552**	**0.408**	0.277	**0.144**	0.42	0.68	0.33	0.00	0.61
ling+pr(5k)	0.546	0.399	0.297†	0.142	0.40	0.68	0.33	0.00	0.58
all/nogr	0.538	0.343	0.273	0.116	0.42	0.68	0.10	0.00	0.52

Table 2: Experiment I: Baselines, specialized linguistic features (*ling*), word/lemma/pos bigrams (*bi*), word pairs (*wp*), CFG productions (*pr*), and combination of linguistic and shallow features, and of all feature groups (*all/nogr*). None of the results show significant improvements over the linguistic features (*ling*) according to McNemar's test.

when system and/or gold standard contain multiple labels and both label sets overlap. As an average over relation types, we can also calculate an average of the F-score over all relations, yielding the *macro-averaged F-score* (MAFS; see Section 7.1).

We also performed statistical significance testing using McNemar's test (Dietterich, 1998; McNemar, 1947) by counting the transitions from wrong labelings (no overlap) to partially or totally correct ones, and vice versa. Under the null assumption, the number of improvements should be similar to the number of disimprovements, and the difference between the numbers of improvements and disimprovements is assumed to be χ^2-distributed.

Because the label distribution is heavily skewed – some relations, such as *Restatement*, are relatively frequent with 140 occurrences, while, e.g., *Contrast* with 26 occurrences, is much less frequent – a classification that is biased towards the more frequent relations will receive higher token-weighted (Dice) scores and lower type-weighted (MAFS) scores, whereas an unbiased system would receive lower Dice and higher macro-averaged F scores.

Table 3 compares the results that are obtained for the full graph (*grA*), a version with all features except information status (*grB*), and finally a minimal version that excludes all semantic features and lemmas (*grC*).

In **Experiment I** (Table 2), we see that the linguistic features perform much better than the shallow features, where in turn bigrams and word pairs reach a higher Dice score than the most-frequent-relation baseline.

We also see that a combination of linguistic and all shallow features (*all/nogr*) performs less well than the best-performing combination of the linguistic features with one of the shallow features (*ling+wp*). For most of the shallow features, there is a threshold for the number of features below which the classification quality deteriorates, and above which the results get slightly worse even

	5 relations		21 relations		Cont	Expn	Temp	Comp	Rept
	Dice	MAFS	Dice	MAFS	F_1	F_1	F_1	F_1	F_1
grA(20k)	0.543	0.364	0.245	0.158	0.38	0.70	0.18	0.00	0.57
grB(20k)	0.552	0.372	0.266	0.167	0.36	0.69	0.20	0.03	0.58
grC(20k)	**0.566‡**	**0.398**	**0.291**	**0.172**	0.38	0.70	0.27	0.06	0.58
ling+grA(20k)	0.580‡	0.387	0.276	0.156	0.37	0.70	0.27	0.00	0.60
ling+grB(20k)	0.577‡	0.396	0.294	0.172	0.36	0.71	0.30	0.00	0.61
ling+grC(20k)	**0.588‡**	**0.408**	**0.307†**	**0.180**	0.37	0.71	0.34	0.00	0.62
allA	0.572‡	0.406	0.304†	0.176	0.42	0.70	0.29	0.00	0.62
allB	0.569‡	0.407	0.297	0.174	0.39	0.70	0.31	0.00	0.62
allC	**0.580‡**	**0.414**	**0.306†**	0.177	0.40	0.71	0.33	0.00	0.63
allA-pr	0.576‡	0.411	0.296	0.175	0.43	0.70	0.33	0.00	0.60
allB-pr	0.582‡	**0.414**	0.293	0.174	0.41	0.70	0.34	0.00	0.62
allC-pr	**0.583‡**	0.413	**0.310†**	**0.178**	0.40	0.71	0.35	0.00	0.62

Table 3: Experiment II: Graph-based representations for discourse segments: All information (*grA*), all without information status (*grB*) and a delexicalized version of the graph representation (*grC*). All graph representations are shown by themselves (*gr*), in combination with linguistic features (*ling+gr*), in combination with linguistic and all shallow features (*all*) as well as with linguistic, bigram and word pair features (*all-pr*). Significant improvements over linguistic features are marked (‡: $p < 0.0001$ according to McNemar's test, †: $p < 0.01$).

though they remain quite stable. In the case of bigrams and grammar productions, this threshold seems to lie around 1000 selected features, while the useful threshold for word pairs seems to lie around 2000 selected features.

None of the improvements from the shallow features is a statistically significant improvement according to the test: even in the case of the word pair features, which give a visible improvement in score, the number of disimprovements is greater than the number of improvements; the number of exact matches stays about the same.

Table 3 presents the results for **Experiment II** concerning the use of graph structures (together with the use of subgraph extraction and feature selection).[9]

The graph structures by themselves already give results that are superior to those using linguistic features. They provide a higher precision on *Expansion* relations, and generally better performance on *Reporting* relations, is the only information source to provide enough information for the identification of *Comparison* relations by themselves. In terms of statistical significance, the improvements at the coarsest level correspond to a highly significant difference in behaviour: from the linguistic features (*ling*) to the graph features (*grA*), there are 321 improvements and 178 disimprovements, yielding a value of $X^2 \approx 40.7$ ($p < 0.0001$). On the finest level (21 relations), the linguistic features and the graph features are not significantly different ($p \geq 0.4$). (Similar things hold for combinations of the graph features and other features — for the combination of all features without CFG productions, the fine-grained level shows 176 improvements and 147 disimprovements, which does not yet reach statistical significance at $p \approx 0.11$).[10]

The second group of rows in Table 3, with combinations of the linguistic features with the shallow information sources and with the graph representation, shows that the graph-based representation works best also when compared to the combinations of linguistic and shallow features in Table 2.

[9] Results differ from previous versions because the subgraph isomorphism detection from the VFLib library (Cordella et al., 2004) turned out to give different results to gSpan's internal matcher. The new results from Tables 3 and 6 directly use the matches output by the gSpan code.

[10] A non-stratified paired t-test over the ten folds is considerably less sensitive; using the paired t-test, the Dice scores of *ling+grB*, *ling+grC*, *allC*, *allB-pr* and *allC-pr* are all significantly higher than *ling* at $p < 0.05$, whereas the test is too weak to detect other differences.

	5 relations		21 relations		Cont	Expn	Temp	Comp	Rept
	Dice	MAFS	Dice	MAFS	F_1	F_1	F_1	F_1	F_1
Restatement	0.474	0.129	0.161	0.014	0.00	0.65	0.00	0.00	0.00
random	0.338	0.233	0.096	0.056	0.06	0.50	0.27	0.21	0.14
SVMperf									
ling only (deg1)	0.476	0.134	0.167	0.025	0.00	0.65	0.00	0.00	0.03
ling only (deg2)	0.540	0.396	0.274	0.127	0.40	0.68	0.32	0.00	0.58
ling+gr(20k)	**0.574**	0.389	0.285	0.161	0.37	0.70	0.28	0.00	0.59
ling+gr(all)	0.506	**0.431**	0.253	**0.181**	0.41	0.66	0.29	0.16	0.63
SVMlight/TK									
ling only (deg1)	0.497	0.406	0.242	0.162	0.37	0.63	0.10	0.31	0.61
ling only (deg2)	0.548	0.435	0.274	0.180	0.40	0.66	0.32	0.32	0.69
ling+gr(PTK)	0.507	0.346	0.211	0.104	0.35	0.64	0.19	0.00	0.54

Table 4: Experiment III: Different methods of classification using the graph structures.

We also see in the first group of rows in Table 3 that the graph representation with the richest set of features performs best in isolation, with a drop in the Dice score when one uses a more impoverished representation. In contrast, the richer representation works less well when it is combined with the linguistic features. In this case, the most impoverished representation performs best.

When all the shallow features are added to the linguistic and graph features, we see a (relatively slight) drop in the Dice score, despite the fact that the F_1 measure for individual relations and the macroaveraged F-score show a slight improvement. Removing the grammar productions from the set of information sources (fourth group of rows in Table 3), we again get a slight improvement. Similar to the group of results combining linguistic and graph features, we see that the parsimonious *grC* graph gives the best combination result (*allC–pr*, including linguistic, word pair, unigram/bigram, and graph features) despite the more informative *grA* giving the best results in isolation.

Looking at individual relations, we see that the identification of rare relations such as *Temporal*, *Comparison*, and *Reporting* is helped by the graph representation (the full system obtains the best MAFS scores of 0.438 and 0.208, for coarse- and fine-grained relations, respectively, against 0.388 and 0.145 for the system without graph information). System variants with graph information also obtain higher coarse-grained Dice scores (0.559–0.581) than the version without graph information (0.552 for *ling+wp* and 0.538 for *all–gr*).

For **Experiment III** (Table 4), we compare the linearization-based approach of extracting subgraphs from the graph representation and treating them as features for the SVM_{perf} classifier with an approach that uses the partial tree kernel to take into account the graph structure.

In the second group of rows of Table 4, we see that feature combination (i.e., forming features from two of the original linguistic features) is essential for the performance of the linguistic features. We also see that feature selection is essential for the performance of the graph-based approach: when taking all 500 000 subgraphs instead of the most informative 20 000 ones, the additional noise drowns even the information from the linguistic features and results in a Dice score of only 0.506 (against 0.540 for the linguistic features by themselves).

In the experiments using the SVMlight/TK classifier, we see parallel tendencies: we see that a linear SVM (deg1) does not perform as well as an SVM classifier using a degree-2 polynomial kernel (deg2). The classification results using the partial tree kernel are very similar to the graph-based approach without feature selection, resulting in a Dice score of 0.507.

In **Experiment IV** (Table 5), we investigate whether the thresholds initially chosen — at most three feature nodes, at least two non-feature edges, at least five occurrences — are in fact sensible, and how sensitive our approach is to choice of the parameter.

It turns out that a small setting of the *minimal support* threshold is in fact beneficial to the performance – it is possible to lower the threshold to two without any adverse effects (but note that

	5 relations		21 relations		Cont	Expn	Temp	Comp	Rept
	Dice	MAFS	Dice	MAFS	F_1	F_1	F_1	F_1	F_1
Restatement	0.474	0.129	0.161	0.014	0.00	0.65	0.00	0.00	0.00
random	0.338	0.233	0.096	0.056	0.06	0.50	0.27	0.21	0.14
ling+grC, defaults	0.588	0.408	0.307	0.180	0.37	0.71	0.34	0.00	0.62
Minimum support									
$f \geq 2$	0.588	0.408	0.303	0.180	0.36	0.70	0.37	0.00	0.61
$f \geq 4$	0.586	0.414	0.304	0.171	0.38	0.71	0.34	0.03	0.61
$f \geq 5$ (default)	0.588	0.408	0.307	0.180	0.37	0.71	0.34	0.00	0.62
$f \geq 8$	0.574	0.408	0.302	0.168	0.40	0.70	0.32	0.00	0.62
$f \geq 40$	0.551	0.392	0.289	0.170	0.37	0.70	0.33	0.00	0.56
Maximum feature edges									
$N \leq 1$	0.560	0.405	0.280	0.179	0.36	0.70	0.33	0.04	0.60
$N \leq 2$	0.582	0.419	0.304	0.174	0.41	0.71	0.33	0.03	0.62
$N \leq 3$ (default)	0.588	0.408	0.307	0.180	0.37	0.71	0.34	0.00	0.62
$N \leq 4$	0.584	0.403	0.303	0.186	0.37	0.71	0.34	0.00	0.60
$N \leq 5$	0.585	0.407	0.306	0.193	0.38	0.70	0.34	0.00	0.60
$N \leq 6$	0.580	0.403	0.305	0.192	0.39	0.70	0.33	0.00	0.60
Minimum backbone nodes									
$M \geq 0$ (default)	0.588	0.408	0.300	0.173	0.37	0.71	0.34	0.00	0.62
$M \geq 1$	0.595	0.416	0.311	0.181	0.37	0.71	0.37	0.00	0.63
$M \geq 2$	0.585	0.408	0.300	0.173	0.36	0.71	0.36	0.00	0.62
$M \geq 3$	0.544	0.361	0.247	0.142	0.34	0.69	0.32	0.00	0.46
$M \geq 4$	0.550	0.396	0.259	0.126	0.38	0.69	0.37	0.03	0.51

Table 5: Experiment IV: Thresholds for subgraph extraction.

the χ^2-based feature selection that only keeps the 20 000 most informative features will also filter out any matches that are too infrequent and/or uninformative).

With the number of feature edges, we see a large improvement between allowing only one feature edge (which yields graphs that consist of mostly backbone nodes with at most one additional feature overall) and allowing two of them. As the number of possible feature edges increases towards the limit imposed on the total graph size for efficiency reasons (7 nodes, which would limit structures to 6 feature nodes and one backbone node), we see that classification performance only decreases very slowly, again with the caveat that feature selection already filters out subgraphs that are altogether uninformative. This can be compared with classification results for polynomial kernels, where the inclusion of second-degree interactions yields large gains, but higher-degree interactions between features bring decreasing returns (as well as noise and sparsity problems).

Looking at the version that only allows one feature edge, we see that the results come closer to the linguistic features, but generally stay above the results for just linguistic features. Hence the graph structure plays a dual role with respect to the linguistic features: the linguistic features contain information about the whole sentence that does not necessarily profit from structural information, whereas the graph features' strength consists both in combining local informations with structural information about the surroundings, as well as in the efficient combination of local pieces of information with each other.

Table 6 shows the results of applying the approach described here on data from the Penn Discourse Treebank (**Experiment V**). In this case, the most informative single feature is found in the context-free productions, which give a slightly higher Dice score for themselves than when combined with other features, and also a higher score than other features in isolation or in combination.

If we compare the English results with those obtained on the smaller German corpus, we see that the English linguistic and graph features do not seem to be sufficient to predict *Temporal* relations,

	5 relations		13 relations		Cont	Expn	Temp	Comp
	Dice	MAFS	Dice	MAFS	F_1	F_1	F_1	F_1
Conjunction	0.523	0.172	0.202	0.026	0.00	0.69	0.00	0.00
random	0.363	0.244	0.168	0.072	0.24	0.52	0.07	0.14
ling only	0.538	0.268	0.290	0.066	0.39	0.69	0.00	0.00
bi(5k)	0.536	0.268	0.320	0.091	0.39	0.68	0.00	0.00
wp(2k)	0.535	0.273	0.258	0.081	0.38	0.69	0.00	0.03
pr(2k)	**0.554**	**0.283**	0.347	0.092	0.44	0.69	0.00	0.00
ling+bi(5k)	0.545	**0.349**	0.339	0.120	0.42	0.69	0.00	0.29
ling+wp(2k)	0.544	0.281	0.323	0.091	0.43	0.69	0.00	0.01
ling+pr(2k)	**0.552**	0.281	0.339	0.093	0.44	0.69	0.00	0.00
gr(20k)	0.536	0.273	0.297	0.090	0.39	0.68	0.00	0.02
ling+gr(20k)	0.544	0.278	0.306	0.094	0.41	0.68	0.00	0.01
all	**0.551**	0.282	0.342	0.107	0.43	0.69	0.00	0.01
all-gr	0.549	0.280	0.349	0.114	0.44	0.69	0.00	0.00
all-bi	**0.552**	0.281	0.339	0.102	0.44	0.69	0.00	0.00
all-wp	0.549	**0.283**	0.338	0.104	0.43	0.69	0.00	0.01
all-pr	0.538	0.275	0.321	0.108	0.41	0.69	0.00	0.01

Table 6: Experiment V: Preliminary results on sections 00–02 of the Penn Treebank/Penn Discourse Treebank.

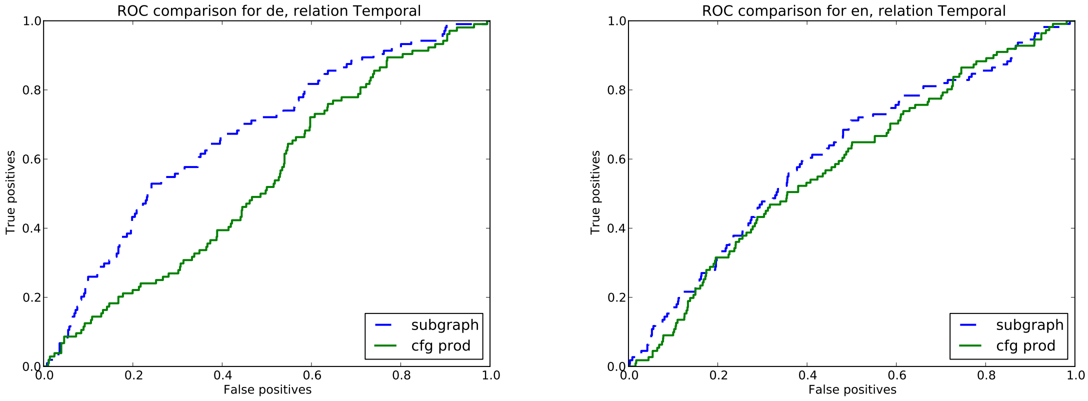

Figure 3: ROC curve comparison for the *Temporal* relation in TüBa-D/Z (left) and PDTB sample (right)

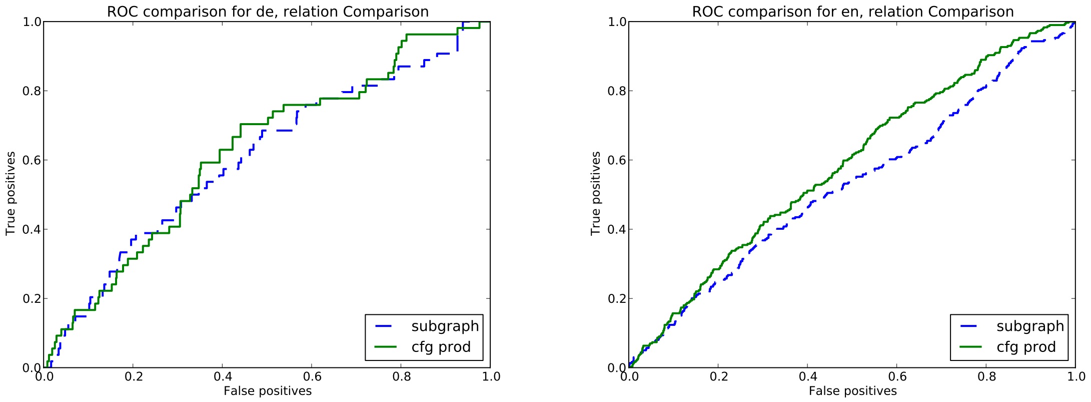

Figure 4: ROC curve comparison for the *Comparison* relation in TüBa-D/Z (left) and PDTB sample (right)

whereas the German linguistic and graph features seem to allow this with some success. For the other relations (*Contingency*, *Expansion*, *Comparison*), we can note that the performance is roughly similar and in both languages appears to be correlated with the relative frequency of those relations, with the most frequent group *Expansion* getting the best prediction results and the relatively less frequent *Contingency* group also reaching an F-score in the middle range (around 0.40–0.43 for German and around 0.44 for English).

In order to have a closer look of the differences with respect to the *Temporal* relation, let us have a look at Figure 3. We see that the German subgraph feature is better than the English subgraph feature when it comes to identifying temporal relations with high precision (lower left corner of left graph, dashed line). At the same time, the grammar productions (continuous line) work less well when it comes to high-precision identification of *Temporal* relations in German than they do in English.

For *Comparison*, another of those relations that are not easy to predict, we see that the difference is less pronounced. While the performance for *Comparison* relations in German is generally higher, we also see that grammar productions and subgraph features stay relatively close together.

7.3 Error analysis

In order to assess the types of errors that occur in the classification, we extracted examples that received high confidence values and therefore provide an indication for a robust true positive, false negative or false positive.

For the **Comparison** relation, the detection of parallel structure is especially important, since *contrast* or *parallel* relations tend to compare two items relative to a common aspect. In the following *Comparison* relation, the parallel structure of a name in the subject and a past-tensed clause (together with an asyndetic conjunction) is detected as an indicator for parallelism:

(8) [Okcuoglu war ein wohlerzogener Bauernsohn und wurde Jurist], [Öcalan wuchs als Kind einer siebenköpfigen armen Bauernfamilie auf].
[*Okcuoglu was a well-bred farmer's son and became a lawyer*], [*Öcalan grew up as the child of a seven-strong poor farmer family*].

The parallel structure is much harder to detect in the following example (which is not identified by the system):

(9) [Eine interne Kostenrechnung geht von 1,4 Milliarden Mark aus], [Mitarbeiter der Kreisverwaltung rechnen eher mit dem Doppelten].
[*An internal cost estimate assumes about 1.4 billion Mark*], [*employees of the local administration rather expect twice that*].

In this case, the common integrator (different parties providing cost estimates) is hidden in the verbs *ausgehen* (here: to assume) and *mit ... rechnen* (to anticipate) is not accessible directly or even through a comparison in GermaNet. The fact that the two contrasting entities (the internal cost estimate and the one from the employees of the local administration) are (specific) indefinites rather than definites, as well as structurally and semantically dissimilar, does not make the task easier; finally, the parallelism between *1,4 Milliarden Mark* (1.4 billion Mark) and *dem Doppelten* (twice that) requires somewhat more complex inference.

Conversely, a vaguely similar temporal and modal structure is not a good indication for parallelism:

(10) [Genmanipulationen lassen sich im nachhinein nur schwer nachweisen.] [Eine Garantie, daß ihre Produkte gentechnikfrei seien, können selbst Ökobauern nicht mehr ohne Gewissensbisse geben.]
[*Genetic manipulations are hard to detect in retrospect.*] [*Not even organic farmers can give a guarantee that their products are free from genetically engineered substances.*]

Here, two verbs *lassen* (realizing an impersonal passive indicating the (im-)possibility of something) and *können* (also indicating possibility) show a roughly similar structure, but simple argument structure is not sufficient for detecting the causally mediated paraphrase between *detecting genetic manipulations* and *give a guarantee that their products are free from genetically engineered substances* that is needed for detecting the *Instance* relation of this example.

In the case of **Contingency** relations, we often find syntactic marking that is a strong indicator of these relations (analogous to *AltLex* relations in the Penn Discourse Treebank, where some indicator is present that does not fulfill the criteria for a discourse connective).

For example, adverbials (even temporal ones) can be a relatively good indicator of a (causal or concessive) contingency relation:

(11) [In den neunziger Jahren, als er wegen neuer Haftstrafen seinen Beruf vorübergehend einstellte,] [versuchte er sich als Publizist und Verlagsgründer.]
[*In the nineties, when he had to leave his profession temporarily,*] [*he tried his hand as a publicist and founding a publishing company.*]

In absence of information about headedness (which is implicit in the relation labels), longer spans of text become more an more intransparent. The following false negative exemplifies this weakness:

(12) [Die Hamburger Staatsanwaltschaft hat gegen den Sprecher der Bundesarbeitsgemeinschaft kritischer PolizistInnen, Strafbefehl beantragt. Wenn das Gericht zustimmt, drohen Wüppesahl zehn Monate Gefängnis.] [Er soll als Fahnder des Dezernats für “Kfz-Schiebereien” des Landeskriminalamts (LKA 234) insgesamt 68 Ermittlungsakten entwendet haben.]
[*The prosecution in Hamburg has requested a penalty order against the speaker of the national working group of critical policepersons. If the court agrees, Wüppesahl would be threatened with ten months of imprisonment*] [*He is said to have stolen 68 investigation files as an investigator of the department for “car trafficking” of the Landeskriminalamt (LKA 234)*].

In this case, the system fails to find the causal relation between the alleged theft and the penalty order.

In some cases, lexical sparsity means that a 'weak' indicator which would give reliable information is misinterpreted by the classifier:

(13) [Neben den ersten 14,9 Prozent von YPF durch einen Kauf vor einem Jahr,] [besitzen die Spanier noch 56 Prozent des zweiten argentinischen Erdölunternehmen Astra.]
[*Besides the first 14.9 percent of YPF through an acquisition a year ago,*] [*the Spaniards own 56 percent of the second Argentinian petrol company Astra.*]

Reporting relations are most often marked through reporting verbs such as *erzählen* (to recount), *ankündigen* (to announce):

(14) ["Prominentenakten sind weggeschlossen",] [erklärt Gerhardt betont nüchtern.]
["*The files of celebrities are shut away*",] [*Gerhard declares in an austere tone.*]

In cases where the discourse relation holds through an anaphoric link, the identification is considerably more difficult:

(15) [Private Unternehmen dürfen die Telefonbücher der Telekom-Tochter DeTeMedien nicht ohne deren Erlaubnis zur Herstellung einer Telefonauskunfts-CD verwenden.] [Das hat der Bundesgerichtshof (BGH) gestern in Karlsruhe entschieden.]
[*Private companies are not allowed to use the telephone books of Telekom daughter company DeTeMedien without its permission to create a telephone directory CD.*] [*This has been decided by the federal court of law (BGH) yesterday in Karlsruhe.*]

Similar to syntactically mediated *Contingency* relations, lexical sparseness results in clause subordinations occurring as false positives for *Reporting* when the respective material is unknown.

In the case of **Temporal** relations, a past/past perfect sequence of tenses can be used successfully, whereas temporal sequences that are understood because of our understanding of actions are missed:

(16) [Da steht Jusef auf.] [Er wendet den Blick von der Wand] ...
[*So Jusef stands up.*][*He diverts his gaze from the wall*] ...

Finally, **Expansion** relations are typical for large spans of text, and constitute the majority of discourse relations for these. In rare cases, such as right dislocation or fragments — often with a relation between main verb and the head noun of such a fragment — the relation between the fragments cannot be inferred with the means of the current system:

(17) [In leicht zeitversetzten Sequenzen hört man disharmonische Tonfolgen und Lautsprecherpfeifen -] [jenes "Feedback" eben, das er sich als Maler erhofft.]
[*In time-shifted sequences one hears disharmonic note sequences and loudspeaker whistling*] - [*just the kind of "feedback" that he was hoping for as a painter.*]

Taking a step back, we should note that there is a subset of implicit discourse relations that is *tractable* in the sense that, even without the context, knowing the content of both discourse relation arguments would put us in a good position to evaluate whether they are both information pertaining to a common question under discussion (as with *Comparison* relations), or whether one clause is the direct or indirect argument to a suitable verb in the other clause (which is the case for some constructions indicating a *Contingency* or *Reporting* relation), or use knowledge about events to estimate the likelihood of a causal or temporal relation. The annotation scheme of TüBa-D/Z makes learning these indicators more straightforward, as, e.g., discourse relations expressing instances of event causation are annotated as (and can be learned as) instances both of a temporal and of a causal connection.

In many cases we see the limits of a purely supervised approach based on treebank data and taxonomic information such as that available from GermaNet: some important lexical clues, such as the nature of reporting verbs, can be learned in a supervised fashion (the head lemmas in the *ling* feature, or the respective information in the graph descriptions), but suffer from sparsity. In some cases, such as the more difficult cases of *Comparison* relations, it would be necessary to judge the similarity of whole phrases, or the event types realized by the clauses, which is a currently unsolved problem where we might however see improvements. In many cases that are already in the reach of automatic classification, we see that capturing the interaction between lexical and structural information, as is possible with the graph structure, is instrumental to improving the achievable results.

8. Conclusion

In this article, we introduced the notion of *feature-node* graphs, which distinguish between so-called *backbone* nodes, which correspond to elements of linguistic structure, and *feature* nodes which contain information about these nodes in a structured-input classification task. We applied an approach that combines a graph-based representation, more detailed linguistic information and shallow surface features to the prediction of unmarked discourse relations in German, with a comparative experiment aimed at understanding the differences with respect to English.

For German, we found that the graph-based representation, when paired with subgraph extraction and supervised feature selection, significantly improves the classification performance over a strong baseline of linguistic features, where shallow features bring a more limited improvement (which does not reach statistical significance). Comparing different variants of the approach of graph construction, we find that, similar to nominal features, noisy or sparse information can hurt classification performance — to counter this, careful construction of the graph, as well as supervised feature selection are effective tools. In a related experiment, we used the graph-based representation together with support vector machines and the partial tree kernel implementation in SVMlight/TK. The results of this experiment, where tree kernels, similarly to subgraph extraction without feature selection, perform relatively poorly, indicate that feature selection is indeed one of the essential ingredients for filtering the broad set of possible subgraphs down to a much narrower set of informative subgraphs.

Finally, we presented preliminary experiments using English data. On the small subset of the Penn Discourse Treebank that we used for comparison, we see a similar general tendency for the most frequent relations to be well-represented in the classification, while rarer relations are less reliable. In contrast to German, we find that the classification of *Temporal* relations reaches somewhat lower performance, and we also see a different relative importance of grammar productions on one hand and subgraph features on the other. Possible causes for this discrepancy include differences in the grouping of discourse relations (where the TüBa-D/Z puts relations in the *Comparison* group whenever they are correlated with structural parallelism — including the *parallel* relation — whereas the PDTB's definition hinges on the more abstract criterion that the author wants "to highlight prominent differences between the two situations"), or in the resources used in the graph construction. A more elaborate graph construction, and possibly more meaningful comparison for English, will be subject of future work.

Acknowledgements The research reported in this article was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of SFB 833 (*Bedeutungskonstitution*). The author is grateful to the three anonymous reviewers for helpful suggestions.

References

Afantenos, Stergos and Nicholas Asher (2010), Testing SDRT's right frontier, *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).*

Arora, Shilpa, Elijah Mayfield, Carolyn Penstein-Rosé, and Eric Nyberg (2010), Sentiment classification using automatically extracted subgraph features, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010).*

Asher, Nicholas (1993), *Reference to Abstract Objects in Discourse*, Kluwer, Dordrecht.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski (2003), Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory, *Current Directions in Discourse and Dialogue*, Kluwer.

Collins, Michael and Nigel Duffy (2002), Convolution kernels for natural language, *Advances in Neural Information Processing Systems 14.*

Cordella, L.P., P. Foggia, C. Sansone, and M. Vento (2004), A (sub)graph isomorphism algorithm for matching large graphs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26** (10), pp. 1367–1372.

Dietterich, Thomas G. (1998), Approximate statistical tests for comparing supervised learning algorithms, *Neural Computation* **7** (10), pp. 1895–1923.

Feng, Vanessa Wei and Graeme Hirst (2012), Text-level discourse parsing with rich linguistic features, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012).*

Gastel, Anna, Sabrina Schulze, Yannick Versley, and Erhard Hinrichs (2011), Annotation of implicit discourse relations in the TüBa-D/Z treebank, *Jahrestagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL 2011).*

Haddow, Barry (2005), *Acquiring a disambiguation model for discourse connectives*, Master's thesis, School of Informatics, University of Edinburgh.

Henrich, Verena and Erhard Hinrichs (2010), GernEdiT - the GermaNet editing tool, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pp. 2228–2235.

Hernault, Hugo, Danushka Bollegala, and Mitsuru Ishizuka (2010), A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010).*

Höhle, Tilman (1986), Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder, *Akten des Siebten Internationalen Germanistenkongresses 1985*, pp. 329–340.

Joachims, Thorsten (2005), A support vector method for multivariate performance measures, *Proceedings of the International Conference on Machine Learning (ICML).*

Kipper, Karin, Hoa Trang Dang, and Martha Palmer (2000), Class-based construction of a verb lexicon, *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI 2000).*

Klenner, Manfred, S. Petrakis, and A. Fahrni (2009), Robust compositional polarity classification, *Recent Advances in Natural Language Processing (RANLP 2009).*

Kudo, Taku, Eisaku Maeda, and Yuji Matsumoto (2004), An application of boosting to graph classification, *Eighteenth Annual Conference on Neural Information Processing Systems (NIPS 2004).*

Lin, Ziheng, Min-Yen Kan, and Hwee Tou Ng (2009), Recognizing implicit discourse relations in the Penn Discourse Treebank, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009).*

Mann, William C. and Sandra A. Thompson (1988), Rhetorical Structure Theory: Toward a functional theory of text organization, *Text* **8**, pp. 243–281.

Marcu, Daniel (2000), The rhetorical parsing of unrestricted texts: A surface-based approach, *Computational Linguistics* **26**, pp. 3.

McNemar, Quinn (1947), Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, pp. 153–157.

Miltsakaki, Eleni, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber (2005), Experiments on sense annotations and sense disambiguation of discourse connectives, *Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005).*

Moschitti, Alessandro (2006), Making tree kernels practical for natural language learning, *Proc. EACL 2006.*

Moschitti, Alessandro and Silvia Quarteroni (2011), Linguistic kernels for answer re-ranking in question answering systems, *Information Processing and Management* **47**, pp. 825–842.

Muller, Philippe, Stergos Afantenos, Pascal Denis, and Nicholas Asher (2012), Constrained decoding for text-level discourse parsing, *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012).*

Park, Joonsuk and Claire Cardie (2012), Improving implicit discourse relation recognition through feature set optimization, *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012)*, pp. 108–112.

Pasch, Renate, Ursula Brauße, Eva Breindl, and Ulrich Hermann Waßner (2003), *Handbuch der deutschen Konnektoren*, Walter de Gruyter, Berlin / New York.

Péry-Woodley, Marie-Paule, Stergos D. Afantenos, Lydia-Mai Ho-Dac, and Nicholas Asher (2011), La ressource AnnoDis, un corpus enrichi d'annotation discursives, *Traitement Automatique des Langues* **52** (3), pp. 71–101.

Pitler, Emily and Ani Nenkova (2009), Using syntax to disambiguate explicit discourse connectives in text, *ACL 2009 short papers.*

Pitler, Emily, Annie Louis, and Ani Nenkova (2009), Automatic sense prediction for implicit discourse relations in text, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009).*

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber (2008), The Penn Discourse Treebank 2.0, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (2010), SentiWS — a publicly available German-language resource for sentiment analysis, *Proceedings of LREC 2010.*

Sanders, Ted J. M., Wilbert P. M. Spooren, and Leo G. M. Noordman (1992), Toward a taxonomy of coherence relations, *Discourse Processes* **15** (1), pp. 1–35.

Severyn, Aliaksei and Alessandro Moschitti (2013), Fast linearization of tree kernels over large-scale data, *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.*

Soricut, Radu and Daniel Marcu (2003), Sentence level discourse parsing using syntactic and lexical information, *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003).*

Sporleder, Caroline and Alex Lascarides (2008), Using automatically labelled examples to classify rhetorical relations: An assessment, *Natural Language Engineering* **14** (3), pp. 369–416.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck (2009), Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z), *Technical report*, Seminar für Sprachwissenschaft, Universität Tübingen.

Versley, Yannick (2006), A constraint-based approach to noun phrase coreference resolution in German newspaper text, *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006).*

Versley, Yannick (2011a), Multilabel tagging of discourse relations in ambiguous temporal connectives, *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011).*

Versley, Yannick (2011b), Towards finer-grained tagging of discourse relations, *Beyond Semantics: Corpus-based investigations of pragmatic and discourse phenomena (Workshop at the annual meeting of the DGfS).*

Versley, Yannick and Anna Gastel (2013), Linguistic tests for discourse relations in the TüBa-D/Z treebank of German, *Dialogue and Discourse* **4** (2), pp. 142–173.

Wolf, Florian and Edward Gibson (2005), Representing discourse coherence: A corpus-based study, *Computational Linguistics* **31** (2), pp. 249–287.

Yan, Xifeng and J. Han (2002), gSpan: Graph-based substructure pattern mining, *Proceedings fo the Second IEEE Conference on Data Mining (ICDM 2002).*